Mapping worldwide language use and multilingualism through public engagement

TESS WOOD & MICHELLE MORRISON

This article appears in:

Endangered languages and the land: Mapping landscapes of multilingualism

Proceedings of the 22nd Annual Conference of the Foundation for Endangered Languages (FEL XXII / 2018)
Vigdís World Language Centre, Reykjavík, 23–25 August 2018

Editors: Sebastian Drude, Nicholas Ostler, Marielle Moser


Cite this article:

First published: December 2018

Link to this article: http://www.elpublishing.org/PID/4008

This article is published under a Creative Commons License CC-BY-NC (Attribution-NonCommercial). The licence permits users to use, reproduce, disseminate or display the article provided that the author is attributed as the original creator and that the reuse is restricted to non-commercial purposes i.e. research or educational use. See http://creativecommons.org/licenses/by-nc/4.0/

Foundation for Endangered Languages: http://www.ogmios.org
EL Publishing: http://www.elpublishing.org
Mapping worldwide language use and multilingualism through public engagement

Tess Wood
University of Maryland Language Science Center
The University of Maryland • 2130 HJ Patterson Hall • College Park, Maryland 20742
[ewood1@umd.edu]

Michelle Morrison
University of Maryland Center for Advanced Study of Language
The University of Maryland • 7005 52nd Ave • College Park, Maryland 20742
[mmorriso@umd.edu]

Abstract
This paper describes a developing initiative at the University of Maryland which aims to produce open-use worldwide geolinguistic data suitable for a broad range of applications. This initiative will integrate data from a range of different sources, including published research, data mining of online resources, and crowdsourcing efforts. Widespread global access to the internet, including the rapid spread of mobile devices in many countries, means that there is enormous potential to gather information through mechanisms as simple as a survey shared on social media, or a poll provided through text messaging. In order to engage diverse contributors, we are developing a range of methods for gathering data. These include brief surveys and more in-depth online tools, as well as extensive networking and outreach plans to reach specific populations with important expertise, such as linguists, translators, humanitarian workers, and community leaders. Our ultimate goal is to develop a broad, encyclopedic resource that serves multiple applications and user groups: researchers, language teachers and learners, K-12 education (across multiple curriculum areas), government, humanitarian responders, communities and speakers of minority and underserved languages, and the general public.

Background
Point to any spot on the globe. What languages are spoken there? By whom? In what contexts? The answers to these questions are almost always complex and hard to find – yet have many valuable applications for research, education, diplomacy, security, humanitarian aid and international development. This paper will describe several innovative aspects of ongoing work in language mapping and geolinguistic data at the University of Maryland (UMD), including developments in crowdsourcing methodology as well as a planned project in collaboration with the NGO Translators Without Borders to develop language maps for humanitarian applications.

The relationships between language and space are complex. Much research within the field of geolinguistics has focused on the creation of worldwide linguistic atlases (e.g., Asher and Moseley 2007), and/or the creation of language maps for a particular geographical area (e.g., van der Merwe and van der Merwe 2006). Geolinguistics has also found a diverse array of applications within the field of sociolinguistics, including the creation of dialect maps and atlases (e.g., Labov, Ash, and Boberg 2006) and the use of maps within perceptual dialectology (Preston 1999). Other geolinguistic projects, such as the World Atlas of Language Structures (Dryer and Haspelmath 2013), focus on the geographical distribution of linguistic features.

There are many inherent challenges in visualizing a particular linguistic variety as belonging to a specific, bounded region in space. Language and place are in many-to-many relationships which are not static as populations move and as patterns of language use shift. Any large-scale language mapping project is faced with the challenge of reconciling data from a variety of sources that may differ significantly from one another and may even provide conflicting information, due to numerous factors, including differences in the time or methods by which the data set was collected, and (more fundamentally), differences in underlying questions and motivations, or in definitions of what precisely constitutes a “language” or a “speaker of a language.” These factors all pose problems both for data accuracy and for visualization of data.

As a result, existing data about language and location are highly variable, often limited in scope, and not easily integrated or made compatible with each other. The World Language Mapping System (WLMS, www.worldgeodatasets.com/language/) is one of the most comprehensive geo-referenced language datasets available, covering the globe. However, WLMS is limited in application by its focus on “linguistic homelands” as well as to the fact that many languages it deems to be widespread are not represented in the dataset (e.g. lingua francas). This focus on first language speaker populations (which is shared by most language atlas projects)
renders the data much less useful for applications, including humanitarian uses, that require a more comprehensive understanding of the languages used in a given area.

There are other geolinguistic resources that amalgamate information from multiple sources within a single interface. For example, the Language and Location — Map Accessibility Project (LL-MAP, www.llmap.org) is a collection of a wide variety of maps. These are primarily language maps, both historical and contemporary, though some other types of maps are also included (e.g., maps of ethnic populations and of migration patterns). LL-MAP presents the maps as digitized versions of the original source; the viewer allows the user to layer multiple maps on top of one another. However, as the primary goal of LL-MAP is to digitize existing maps, the maps cannot be used to answer the same questions or even the same types of questions. Instead, the primary benefit of LL-MAP is its ability to overlay maps visualizing different types of information on top of one another. Thus, these maps cannot be expected to present a seamless or consistent depiction of language use around the world.

Some country-specific maps attempt to provide a more detailed view of language distribution, using data from censuses or similar countrywide datasets. For example, the Modern Language Association’s Language Map (apps.mla.org/map_main) maps the location of more than 300 languages within the United States, using data from the American Community Survey (ACS), conducted by the US Census Bureau. However, the utility of the MLA’s Language Map is limited by the constraints of ACS data (for example, African languages have been grouped together).

As Mackey (1988) observes, the challenge of language mapping is much more than simply mapping the geographic location within which a significant population of a language’s speakers may be found. Instead, it is also critical to study what Mackey terms the “geography of language functions.” This includes understanding the domains in which a language is used — for example, is its use restricted primarily to the home domain? Or is it also used as a medium of education? Is it the official language of the country? These and other questions are absolutely critical for geolinguistic data collection efforts that are designed with real-world applications in mind.

**Previous geolinguistics work at UMD**

Researchers at UMD’s Language Science Center and Center for Advanced Study of Language have been working on several related geolinguistics projects over the last five years. One major collaborative effort, known as Langscape, is a public-facing online language map intended to connect and make accessible diverse information about languages worldwide.

Until recently, Langscape used the World Language Mapping System data for its primary language map layer. As noted above, this is probably the most comprehensive such dataset available, though with several limitations. Using WLMS in combination with other sources of language information, researchers at UMD developed various mechanisms to allow users to access and explore geolinguistic and associated data. These include creating means to overlay multiple map layers and integrate explanatory text and other resources with the map in order to provide a more detailed and nuanced view of the distribution and use of the world’s languages.

As of early 2018, UMD ended its license for WLMS in order to pursue development of more detailed and flexible geolinguistic data and tools that would be open-use and widely accessible for diverse users.

**Gathering new data: Building an open, worldwide dataset**

The ultimate goal of the Langscape project is to develop a broad, encyclopedic resource that can serve multiple applications and user communities: researchers, language teachers and learners, K-12 education (across multiple curriculum areas), government, humanitarian responders, communities and speakers of minority and underserved languages, and the general public. To serve these diverse audiences, it is crucial to provide worldwide language mapping resources that are:

- Inclusive of details of language use and status as well as location;
- Integratable with other types of language information;
- Easily updatable as situations and demographics change;
- Transparent about the data sources reflected in the maps; and
- Freely available for download, use and adaptation by other organizations.

Data will come from a variety of sources, none of which alone would be sufficient to create the kind of flexible resources envisioned. For example, census data in many countries may be quite accurate but lack linguistic detail (e.g., languages may be known by multiple names and thus speakers artificially separated, or may be grouped together into large catch-all categories such as “Indigenous Languages”), and these details may be supplied by other complementary data sources. Some examples of data sources include:

- Published research on the distribution of linguistic and ethnic groups in various countries and regions;
- Existing language maps;
- Data-mining: social media and other online media;
- Country-level and local census data;
- Linguists, anthropologists, and human geographers with regional expertise;
- Translators, language teachers and international professionals with language expertise connected...
with aid and development activities or with other organizations working in the regions of interest; and

- General public via crowdsourcing and related methods.

Gathering and integrating different types of data to build (and keep current) language maps for the entire world is no small feat, and is necessarily dependent on the engagement of a community of experts and non-experts alike in creating and contributing to the resource.

One challenge projects with a global focus often face is that if data is collected gradually for the entire world, there is a long lag time until enough data is available (for any part of the world) to be usable. We plan to take a different approach, gathering data for specific applications, typically defined by a specific geographical region. This allows work to be focused on the most appropriate and valuable sources for the area or problem in question. For example, the availability of academic research, census data and social media data (and choice of platforms) will vary significantly from country to country. Intensive efforts focused on a specific country or area should yield enough data to enable immediate applications (as outlined for the pilot project described below), but still allow data to continue to be added & updated, resulting in a dataset that increases in both breadth and depth over time.

One of the more novel data collection strategies we propose is crowdsourcing of language use data through engaging both expert and non-expert contributors. This will be the focus of the next several sections, including design of crowdsourcing tools and planned outreach efforts to engage participants (taking into account factors such as motivation to participate and access to technology).

**Crowdsourcing**

Crowdsourcing language use information is a potentially transformative method for building worldwide coverage and compiling detailed and accurate data for linguistically complex regions. Widespread access to the internet, including the rapid spread of mobile devices in many countries, means that there is enormous potential to gather information through mechanisms ranging from surveys shared on social media or through text messaging to dedicated apps and interactive websites.

Crowdsourcing and citizen science have proven popular and effective in many fields for gathering large amounts of information, such as documenting distribution of plant species or community mapping following a natural disaster, and for performing “human intelligence tasks” (a term used by Amazon for their Mechanical Turk platform), like identifying characters or objects in an image. Citizen science specifically involves engaging a non-expert volunteer community in gathering, labeling or analyzing data for scientific purposes – for example, identifying animals, constellations, or written characters in an image. The site Zooniverse ([www.zooniverse.org](http://www.zooniverse.org)) provides many examples of such projects in a variety of fields which have been highly successful.

In order to effectively engage participants, crowdsourcing and citizen science projects need to take into account the type of task or information required, the target contributor audience, and the likely motivations of contributors. Depending on these factors, a project might be presented as a game, provide payment of some kind, or depend on an altruistic desire to help or willingness to contribute to a resource (Sabou, Bontcheva & Scharl, 2012). Citizen science projects are volunteer-based and most often have the following characteristics:

1. **Task**: Requires human perceptual and cognitive abilities but little prior knowledge or training. E.g. identifying animals in an image using a provided field guide.
2. **Audience**: Anyone able to read the instructions and complete the task. Audience need not be diverse or have experience.
3. **Motivation**: Contributing to scientific knowledge or resources; learning about research; enjoyable challenge and/or site-specific reward (e.g. badges, points).

Gathering data on language use in different contexts and locations poses some specific challenges for applying this kind of model. Valuable linguistic information is widely distributed (almost all adults know at least one language) but also highly specialized (most adults have experience with relatively few languages and locations). These constraints make the audience of paramount importance in guiding the design of data collection.

Unlike a typical citizen science project, the contributors engaged in this type of project have to be unusually diverse and will differ in:

- Geographical location;
- Access to information and technology, and preferences for specific technologies;
- Motivations and barriers to contributing language information;
- Language expertise (e.g. linguists, language teachers, general public);
- Demographic and linguistic background.

How can we engage a diverse enough community of participants able to provide the breadth of information we hope to gather? In order to address this challenge, we are developing several methods for gathering crowdsourced data. These include brief surveys and more in-depth online tools, as well as extensive networking and outreach plans to reach specific populations with important expertise, such as linguists, translators, humanitarian workers, and community leaders.

**Design considerations for crowdsourcing**

In order to design effective crowdsourcing, we must observe the type of data required and the tasks which will elicit it, the contributors who can complete those tasks, and the likely motivations for their participation.

**Data**: There are several types of information which could be requested from contributors.
Individual reported data: a person’s description of their own language use.

Observed data: language use a person witnesses in a community, perhaps referencing concrete examples within a specific period of time.

Generalized or impressionistic data: a person’s knowledge and beliefs about language use in a geographical area based on the sum of their experience.

Contributors: Individual and observed data can in principle be provided by anyone who knows and uses a language and interacts with others in a real-world (geographical) community of some type. However, we suggest that generalized or impressionistic data (e.g. which might be provided by drawing a polygon to indicate the spread of a particular language community) is more likely to be provided by those with specific experience attending to language use patterns, e.g. linguists, translators, language teachers, community language activists, etc. We will therefore make a distinction between “expert” and “non-expert” contributors, based mainly on their type of contribution.¹ (Experts will be self-identified but most likely have the kinds of expertise listed above). (See Table 1, next page.)

For non-expert contributors, requesting individual data has the benefit of calling on people’s detailed knowledge of their own behaviour. However, collecting only individual data has drawbacks. First, a large number of individual data points must be collected to build up a picture of language use in a community. Second, people’s beliefs about their own behaviour can be subject to biases (which might be partly compensated for by taking into account people’s reports of others’ language use). And third, asking only about an individual’s own behaviour fails to capture their knowledge of a community. Local community members and regular or long-term visitors may have a lot of expertise about language use in an area – and may well be able to answer questions about the contexts in which they have heard or seen a language used recently, for example. We propose that asking only about an individual’s own behaviour may also impact their motivation (since it tends to place them more in the role of research subject than co-creator of scientific knowledge).

Motivation. Motivations for participating in crowdsourcing of language use information can be expected to differ between groups of contributors, particularly between “expert” and “non-expert” contributors. Effective design of surveys and tools depends on careful consideration of contributors’ knowledge and motivation to participate. We have evaluated a number of motivations that might lead people to participate in tasks related to language use in the communities they are connected to.

1. Fun – playing a game;
2. Self-knowledge – getting feedback or learning something about oneself;
3. Self-expression – wanting to be heard/sharing knowledge or opinions;
4. Recognition – gaining professional or social standing or recognition of expertise (e.g. others using and citing their contributions);
5. Raising awareness – adding information about a language or community to increase awareness of its significance;
6. Making a contribution – adding information because it will help to build a resource one values, or more broadly because it will be useful for research or education.

An interesting comparison to the collection of geolinguistic data can be found in projects which depend on Volunteered Geographic Information (VGI) such as OpenStreetMap (openstreetmap.org), or the use of the Ushahidi data collection platform (ushahidi.com) for mapping in Haiti following the 2010 earthquake (Norheim-Hagtun and Meier 2010). These examples are particularly relevant because (unlike many citizen science projects) they depend on a geographically diverse – or geographically specific, in the case of Haiti – contributor base, and enable both experts and non-experts to contribute local knowledge. Budhathoki & Haythornthwaite (2012) examined the motivations for contributions to OpenStreetMap and conclude:

Positive and important motivators were found that accorded with ideas of the “personal but shared need” associated with contribution to open-source projects, co-orientation to open-source and geographic knowledge, and attention to participation in and by the community. (Budhathoki & Haythornthwaite 2012:548)

In other words, ideological motivations were primary: a belief in the importance of the project and its goals, and the desire to help accomplish them.

¹ We recognize that many experts will also be community members, e.g. community members who are linguists, translators, and language educators. The explicit mention of community members in the “non-expert” category is intended to reflect the contributions of those non-expert participants who nevertheless have valuable local and linguistic knowledge.
The study also revealed some specific differences between “serious” and “casual” mappers (roughly parallel to our expert/non-expert distinction):

Differences in motivation between serious and casual mappers showed that serious mappers were more oriented to community, learning, local knowledge, and career motivations (although the latter motivation is low in general), and casual mappers were more oriented to general principles of free availability of mapping data (Budhathoki & Haythornthwaite, 2012:548).

While OpenStreetMap does not exploit some of our proposed non-expert motivations such as fun or self-knowledge, the project nevertheless has been relatively successful in attracting a diverse contributor group to date of over four million people. Correspondingly, we anticipate that our project’s focus on open-use data and the applications of the data to research, education and humanitarian applications, will form the main basis for contributors’ motivation to participate.

With these distinctions in contribution type and motivations in mind, we are currently working on designing crowdsourcing interfaces for expert and non-expert contributors. For both groups, the goal is to focus the contributor’s experience at an appropriate level of detail, provide tools that enable them to effectively share their knowledge, feel confident in the information they are contributing, and be able to see the value of their contribution to the outcomes which are important to them.

### Expert data collection

A team of Human-Computer Interaction (HCI) students at Carnegie Mellon University worked with us in Spring 2018 on initial design of a tool for experts to contribute data about language use (Guerrero et al., 2018). The expert tool is designed for users who feel comfortable identifying approximate regions and patterns of language use on a map, without necessarily making reference to specific examples or to their own personal language use.

Experts’ self-selection is also related to their motivation. We anticipate that contributors who identify as experts will be motivated to contribute to a resource which will be valuable to themselves, their field or profession, or a population they engage with. For example, a teacher or interpreter of an underserved minority language might be motivated to raise awareness of the language and needs of the community they work with; a linguist working on several endangered languages might be motivated to share their expertise on those languages and be able to combine their own knowledge with that of others to advance research and scholarship in the area.

Our student design team conducted user interviews and testing with a number of linguists and language experts at US universities to identify what kind of tools and questions enabled experts to contribute their knowledge most effectively. They found clear preferences among their test users for a flexible dashboard of tools which allowed them to create and edit polygons and associate specific types of information with a polygon, as well as for an interaction which began with the selection of a specific language, then allowed them to provide location information and optionally to add more detailed usage information.

One important question relates to the way in which assumptions about motivation and knowledge shape the tool. Features considered important for this user group include:

1. Ability to draw polygons, with options to add more or less detail to the polygons (e.g. embedding smaller polygons within larger polygons to indicate more specific usage contexts);
2. Options to create an account to retrieve one’s own contributions and manipulate them in specific ways;
3. Option to identify oneself and associate individual information with the data – e.g. link to Google Scholar to describe expertise and increase confidence in the data; ability to have contributions be citable;

<table>
<thead>
<tr>
<th>Expected background</th>
<th>Motivations (non-financial)</th>
<th>Type of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguists, translators, language teachers, etc.</td>
<td>1. Recognition (e.g. being able to cite one’s contribution); 2. Raising awareness about a language or community; 3. Making a contribution (to build a resource one values).</td>
<td>Generalized and/or observed</td>
</tr>
<tr>
<td>Community members, long-term visitors/travellers</td>
<td>1. Fun; 2. Self-knowledge, getting feedback 3. Self-expression, sharing knowledge or opinions; 4. Making a contribution (to research or education more broadly).</td>
<td>Observed and/or individual</td>
</tr>
</tbody>
</table>

Table 1: Expert and non-expert contributors

---

2 Current participation numbers can be found at: [www.openstreetmap.org/stats/data_stats.html](http://www.openstreetmap.org/stats/data_stats.html)
4. Design is primarily focused on desktop/web-based deployment (i.e. the ability to make quick, mobile contributions is less likely to be important).

The expert tool enables users to freely draw polygons on the map to outline areas in which particular languages are spoken, and to further identify contexts in which the languages are used – either associated with the entire polygon or with smaller regions within it.

The expert tool is currently in a prototype phase, and will undergo additional development and testing.

Non-expert data collection

Having carefully considered interaction preferences and motivation for experts, we recognized that a number of characteristics would need to be significantly different for non-expert contributors. The expert tool is designed for the contribution of generalized data, abstracting away from specific usage contexts, and using drawing tools to create polygons. Initial user testing confirmed that these tasks would be challenging and potentially intimidating for non-expert contributors, and would likely limit their ability to express the valuable knowledge they possess. Our priority was to find ways to ensure contributors can provide geographic information without imposing demands that did not match their expertise, while maximizing the level of detail they are able to provide about use contexts.

For non-expert users, we also anticipate a much wider range of devices (including mobile and older feature phones) being used to provide information, so it is important for the design to be adaptable to different platforms. Simplifying the input of geographic information and maximizing the use of text-based survey questions will aid adaptability across platforms.

Users drop a pin to identify a location of interest (the pin will be resolved to the closest town or city), and they are then able to list languages they plan to add in that area. For each language, rather than drawing a polygon, they will have the option to indicate that the language is used: (a) Nationwide; (b) Across a state or region; or (c) In a local area which they will identify. The selected area is created by highlighting grid squares (which can be done easily on a touch screen), rather than drawing a precise boundary.

While this simplified map interaction will work for both mobile and desktop browsers, we would like to be able to adapt the tool to also gather data using text-only interactions. In such cases, location data will be constrained but can be provided by identifying the closest city, for example.

Questions focus on language use in a set of commonly experienced domains: home, schools, local community (e.g. grocery stores or markets, community meetings), news media, social media, and government functions. Contributors are asked about the languages they have recently encountered or are aware of in these settings. The goal of framing the questions in this way is to invite
contributors to reflect on their specific experiences, rather than providing generalized impressions.

To keep the interaction brief, the required questions are minimal, but offer options to answer additional questions on a particular topic if desired.

**Pilot project: Crowdsourced language mapping for humanitarian needs**

As outlined earlier, rather than trying to gather data for the entire world at once our approach is to design specific, geographically or application-focused efforts which will enable the creation of smaller, intermediate data sets with important short-term applications and benefits. We are currently designing a pilot project in collaboration with the NGO Translators Without Borders (translatorswithoutborders.org) which aims to collect language mapping data for several countries (Nigeria, Syria, Bangladesh and Ethiopia) where language mapping can help to address urgent humanitarian needs. These humanitarian use case scenarios can bring into focus some of the most complex linguistic situations that worldwide language mapping will need to address.

The need for current, accurate language mapping for humanitarian response is clear. The current lack of available data about language use makes it difficult for organizations to communicate effectively with populations in urgent situations. For example, the right interpreters may not be available to assist refugees, or broadcast messages with urgent crisis information may fail to reach affected groups who are unable to understand the languages used. Printed materials with important preparation information may primarily reach educated men who can read particular languages, proving ineffective at targeting populations of women, children, or disabled people who are especially vulnerable.

**Conclusions and future directions**

We believe crowdsourcing and engagement with the public provides the potential for collection and curation of a detailed dataset that can inform the development of a robust geolinguistic tool. By capitalizing on the knowledge and insights of a wide user community, this methodology has the potential to build a detailed and nuanced picture of language use in communities around the world.

This crowdsourcing initiative represents one piece of a larger vision, whose goal is the development of a comprehensive, spatially explicit, open-source language mapping toolkit. Our goals are to develop a system that is highly flexible, both in terms of the user communities and needs it serves as well as in terms of the types of data by which the system is informed. Crowdsourcing is an important piece of this endeavor; we also plan to incorporate other data types into our interface, including academic literature, census data, image data (e.g., photos of local street signs). The dataset will be open source and freely downloadable; such a tool has the potential to be transformative for users in a wide variety of applications, including research, humanitarian relief, education, diplomacy, security, and international development.

**Acknowledgments**

The project described is based on joint work with: Colin Phillips, Amalia Gnanadesikan, and Paul England at UMD; and with Eric DeLuca at Translators Without Borders. We are also grateful to the following current and former student researchers for their contributions to this project: Phoebe Cao, Emily Grabowski, Maggie Kandel, Kevin Lu, Barbara Vasilchenko, Hieu Vo.

**References**


